Neurocomputing 270 (2017) 159–169

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Distributed content filtering algorithm based on data label and policy expression in active distribution networks

Song Deng[a,*], Dong Yue[a], Aihua Zhou[b], Xiong Fu[c], Lechan Yang[d], Yu Xue[e]

[a] Institute of Advanced Technology, Nanjing University Post & Telecommunication, 210003, Nanjing, China
[b] Global Energy Interconnection Research Institute, Beijing, 102209, China
[c] School of Computer, Nanjing University Post & Telecommunication, Nanjing, 210003, China
[d] International Institute for Earth System Science, Nanjing University, Nanjing, 210093, China
[e] Nanjing University of Information Science & Technology, 210044, Nanjing, China

ARTICLE INFO

Article history:
Received 25 May 2016
Revised 28 February 2017
Accepted 23 March 2017
Available online 16 June 2017

Keywords:
Active distribution networks
Data label
Policy expression
Content filtering

ABSTRACT

With the development of active distribution networks, data transmission is facing a severe security challenge. Secure data transmission is crucial for the real-time and exact control of active distribution networks. However, traditional data encryption methods have difficulty with the real-time control and mass data transmission of the active distribution networks. Additionally, content filtering based on text classification has a strong dependence on the size and type of data. To solve these problems, this paper proposes a novel distributed content filtering algorithm based on data labeling and policy expression (DCF-DLPE). In DCF-DLPE, we design a secure private protocol with data labeling and build a policy rule expression. Four representative datasets are used to evaluate the performance of the proposed algorithm. The comparative results show that for the larger dataset, DCF-DLPE outperforms the DES, AES (256-bit) and Blowfish encryption methods in the average time-consumption. Experimental results also show that compared with text classification algorithms, DCF-DLPE has a clear advantage in terms of filtering accuracy, sensitivity and precision. It is more important that, compared with text classification algorithms, performance of the DCF-DLPE algorithm is independent of the size and type of the dataset.

1. Introduction

As a typical resource-constrained distributed networked control system [1–3], secure and efficient data transmission in the active distribution network is crucial for the safe and stable operation of the system [4,5]. With extensive applications of wireless communication, the Internet and other advanced information communication technologies in active distribution networks, the active power distribution networks are facing the threat of viruses, Trojans and hacker attacks from the Internet. At the same time, malicious attacks from the Intranet may also cause the networked control system of active distribution networks to collapse. The existing security protection scheme of the power secondary system does not consider security problems between the Internet and Intranet in active distribution networks. Especially with the development of strong smart grids, there are more complex access environments, flexible access modes (such as GPRS, WiFi and fiber optic communication), a huge number of intelligent access terminals (such as various types of collection terminals and monitoring terminals) and dynamic and distributed mass data. How to ensure the security of data transmission in active distribution networks and prevent leakage has become a research highlight and a crucial issue that needs to be addressed in the information security protection of active distribution networks. Data encryption is a common method to address leakage of sensitive business data and injection of malicious information in the process of data transmission and interaction. Cryptography is normally used to protect data from unauthorized access. Encryption tools and algorithms have reached a level of maturity where it is very difficult or almost impossible to decrypt data without the right decryption key [6–8]. Data encryption might ensure the secrecy of the plaintext but not the cipher text. This might lead to various types of attacks such as cipher text, known plaintext and chosen plaintext attacks [9]. These encryption methods address the data in transit. Moreover, some encryption approaches are used to prevent sensitive data leakage for data in use and at rest [10]. We know that the encryption algorithms implemented on a resource-constrained

* Corresponding author.
E-mail addresses: dengsong@njupt.edu.cn, ds16090311@163.com (S. Deng), yued@njupt.edu.cn (D. Yue), zhouaihua@geiri.sgcc.com.cn (A. Zhou), fux@njupt.edu.cn (X. Fu), yanglechan@163.com (L. Yang), xueyu_123@nuaa.edu.cn (Y. Xue).

http://dx.doi.org/10.1016/j.neucom.2017.03.087
0925-2312/

device have difficulty in achieving the ideal system performance. Nie et al. [11] gave a comprehensive review for the methods of power evaluation to decrease power consumption. To maintain the confidentiality and authenticity of the data exchanged between the trusted authority and the vehicular ad hoc networks nodes, dual authentication and key management mechanisms for secure data transmission were designed to provide a variety of online premium services to customers through vehicular ad hoc networks [12]. With the continuous enrichment of data acquisition methods in active distribution networks, data is increasingly collected and stored on the cloud servers. To prevent leakage and potential attacks of big data, some approaches were proposed. Hsu et al. [13] proposed efficient and secure group key transfer based on secret sharing over big data to resist leakage and potential attacks, which also significantly reduces the time-consumption of the system. Seo et al. [14] present a mediated certificate-less encryption scheme without pairing operations and applied the scheme to solve secure sharing of sensitive information in public clouds. To address data security, privacy, confidentiality, integrity and authentication in the cloud, Potey et al. [15] proposed a data encryption scheme on the cloud by using a fully holomorphic encryption. Ramachandran et al. [16] noted that encrypting the data in transit is probably the most secure technique for data in transit. However, for distributed real-time control systems in active distribution networks, massive data transmission by using encryption can ensure security, but the transmission delay and the possibility of decryption failure caused by massive data encryption are bound to produce immeasurable losses for real-time control of active distribution networks. For the various kinds of possible cyber-attacks in distribution automation networks, Lim et al. [4] proposed an effective security enhancement protocol and key deployment mechanism. To prevent malicious intelligent electronic devices from accessing intelligent distribution networks, Sun et al. [17] present an intelligent distribution network access control scheme based on an identity cryptosystem. In addition, to solve the trust problem of cloud server providers, Yu et al. [18] present an improved remote data possession check to fix the security flaws of the original protocol.

To better protect business sensitive data from leaking and injecting malicious information, compared with encryption and other passive protections, intelligent content filtering is a proactive security measure. Most of the existing data filtering algorithms are carried out from the view of classification. According to annotated training data sets, the algorithm finds the relationship between data attribution and category, and then determines automatically the classification of the unmarked data by the found relationship. Wang et al. [19] present classification in networked data based on the probability generative model. This method computed the values of latent variables based on Gibbs sampling by building a new probability generative model for the network to obtain the classes of the nodes. Sami et al. [20] proposed a fast and robust sparse approach based on a few labeled samples for hyperspectral data classification. Nouaouria et al. [21] present an improved global-best particle swarm optimization algorithm based on a particle-position update and an interpretation mechanism. Experimental results had shown that the algorithm has higher accuracy and efficiency of data classification for standard datasets. Lin et al. [22] proposed a new method of similarity measure for text classification. A novel classification framework based on fuzzy formal concept analysis is provided to solve the effect of ambiguous words on text classification [23]. Lee et al. [24]proposed a multi-label text categorization based on fuzzy relevance clustering. To improve the efficiency of mass data classification, Triguero et al. [25] proposed a novel distributed partitioning for prototype reduction based on MapReduce. Tang et al. [26] proposed a Bayesian classification approach for automatic text categorization using class-specific features. Wang et al. [27] proposed a unified framework to expand short texts based on word embedding clustering and convolutional neural networks for improving short text classification. However, these content filtering algorithms are only suitable for real-time protection of text files, and cannot protect the security of structured and unstructured sensitive data [28] such as videos, pictures, documents and databases in active distribution networks systems. Therefore, combined with the characteristics of distributed data transmission, this paper proposes a distributed content filtering algorithm based on data label and policy expression in active distribution networks (DCF-DLPE). DCF-DLPE can filter sensitive data through the combination of data label and policy expression to prevent data leakage in the process of network transmission. This paper focuses on proposing a novel data content filtering algorithm for any data type. The contributions of this paper are as follows.

(1) Combined with the application scenario of active distribution networks, the leakage of data storage, transmission and use are analyzed, and a distributed content filtering framework based on data label and policy expression is proposed.
(2) This paper designs a novel secure private protocol with data label to protect the security of data transmission and build a corresponding policy expression.
(3) Based on the actual requirements of distributed data acquisition and control in active distribution networks, a distributed content filtering algorithm based on data label and policy expression is presented.
(4) Compared with the traditional data encryption and text classification algorithms, the proposed algorithm compares and analyzes processing time, performance, filtering accuracy, sensitivity and precision.

The remainder of the paper is organized as follows. Section 2 introduces the distributed content filtering framework based on data label and policy expression. Section 3 introduces a novel secure private protocol based on data label and designs a policy rule expression. Section 4 gives a description of the distributed content filtering algorithm based on data label and policy expression. Section 5 presents some experiments and performance analysis. Conclusions are given in Section 6.

2. Distributed content filtering framework based on data labeling and policy expression

With the extensive access of multiple distributed equipment and component units such as large-scale distributed energy, energy storage, and flexible load, the large amount of operation data in active distribution networks is collected and transmitted to monitoring units by wireless public networks. Then, the data is transmitted from monitoring units to the master control system for analysis and decision by wireless public networks or optical fiber communication. Meanwhile, each module in the master control system interacts by way of optical fiber communication. The structure is shown in Fig. 1. Compared with traditional distribution networks, the active distribution network has a wide range of data collection, a large amount of data, and high real-time accuracy of data analysis. At the same time, the access and plug-and-play of a large number of distributed generations make energy and information interact frequently. It places higher requirements for the security of data transmission in the active distribution networks.

From Fig. 1, we can see that a large amount of state data is transmitted by wireless public networks (such as GPRS, WiFi, ZigBee, etc.) from a remote terminal unit to a monitoring unit and from a monitoring unit to the master station system in an active distribution network. Therefore, in the process of transmission, state data are easily stolen, attacked or tampered with, which leads to leakage. However, frequent encryption and decryption will inevitably lead to the increase of data transmission and processing